8 The Bootstrap

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals. Let T_n be a *statistic*, that is, a function of the data (such as the sample mean) and suppose we want to know $\text{var}(T_n)$. The bootstrap idea has two parts. First, note that the quantity of interest, $\text{var}(T_n)$, is a functional of F: if you knew F you could (at least in principle) compute the variance of T_n. To emphasize this fact, we write $\text{var}_F(T_n)$. Since we don't know F we use a plug-in estimator of the variance, namely, $\text{var}_{\widehat{F}_n}(T_n)$. The second step is to notice that $\text{var}_{\widehat{F}_n}(T_n)$ can be approximated by simulation. This is a good time for an aside on simulation.

SIMULATION. *Let G be a distribution and let $Y_1, \ldots, Y_B$ be iid values drawn from G. By the law of large numbers, $B^{-1}\sum_{j=1}^B Y_j$ converges in probability to $\int y dG(y) = EY$. So we can use $B^{-1}\sum_{j=1}^B Y_j$ as an estimate of $E(Y)$. In a simulation, we can make B as large as we like in which case, the difference between $B^{-1}\sum_{j=1}^B Y_j$ and $E(Y)$ is negligible. Similarly, the sample variance $\sum_j (Y_j - \overline{Y})^2/B$ of $Y_1, \ldots, Y_B$ estimates $\sigma^2 = \text{Var}(Y)$.*

Now back to the bootstrap. We draw B samples of size n from $\widehat{F}_n$, where B is large. For each sample we compute the statistic giving $T^*_{n,1}, T^*_{n,2}, \ldots, T^*_{n,B}$. Finally we compute the sample variance $\text{var}_{\text{boot}}(T_n)$ of $T^*_{n,1}, \ldots, T^*_{n,B}$ to approximate $\text{var}_{\widehat{F}_n}(T_n)$. The bootstrap standard error of T_n is $\text{se}_{\text{boot}}(T_n) = \sqrt{\text{var}_{\text{boot}}(T_n)}$. Here is a summary of the steps:

(1) Draw B bootstrap samples:

$$\begin{aligned}
\text{sample } 1 &= (X^*_{1,1}, \ldots, X^*_{1,n}) \sim \widehat{F}_n \\
\text{sample } 2 &= (X^*_{2,1}, \ldots, X^*_{2,n}) \sim \widehat{F}_n \\
&\vdots \\
\text{sample } B &= (X^*_{B,1}, \ldots, X^*_{B,n}) \sim \widehat{F}_n
\end{aligned}$$

(2) Compute the statistics

$$T^*_{n,1} = T(\text{sample } 1)$$

$$
\begin{aligned}
T^*_{n,2} &= T(\text{sample } 2) \\
&\vdots \\
T^*_{n,B} &= T(\text{sample } B).
\end{aligned}
$$

(3) Approximate the variance $\mathrm{Var}(T_n)$ with the sample variance of $T^*_{n,1}, \ldots, T^*_{n,B}$:

$$
v_{\text{boot}} \equiv \widehat{\mathrm{Var}}(T) = \frac{1}{B}\sum_{b=1}^{B}\left(T^*_{n,b} - \frac{1}{B}\sum_{r=1}^{B} T^*_{n,b}\right)^2.
$$

(4) The bootstrap estimate of the standard error is $\widehat{se} = \sqrt{v_{boot}}$.

We are using two approximations. First, we estimate $\mathrm{var}_F(T_n)$ by $\mathrm{var}_{\widehat{F}_n}(T_n)$, then we approximate $\mathrm{var}_{\widehat{F}_n}(T_n)$ by $\mathrm{var}_{\text{boot}}(T_n)$. Since we can make B very large, we expect that $\mathrm{var}_{\widehat{F}_n}(T_n) \approx \mathrm{var}_{\text{boot}}(T_n)$. The main source of error is in approximating $\mathrm{var}_F(T_n)$ by $\mathrm{var}_{\widehat{F}_n}(T_n)$. To summarize:

$$
\mathrm{var}_F(T_n) \overbrace{\approx}^{\text{not so small}} \mathrm{var}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{\text{small}} v_{boot}.
$$

How do we simulate from $\widehat{F}_n$? Since $\widehat{F}_n$ gives probability $1/n$ to each data point, drawing n points from $\widehat{F}_n$ **is the same as drawing a sample of size n with replacement from the original data.** This is why bootstrapping is sometimes called resampling the data. Here is an example in R:

```
n <- 100
x <- rnorm(n)                               ### create some data
theta.hat <- median(x)                      ### suppose T(F) = the median
B <- 1000
theta.boot <- rep(0,B)
for(i in 1:B){
     xstar <- sample(x,size=n,replace=T)   ### draw a bootstrap sample
     theta.boot[i] <- median(xstar)        ### compute the statistic
     }
var.boot <- var(theta.boot)
se <- sqrt(var.boot)
print(se)
```

The bootstrap can also be used to estimate bias:

$$b_{\text{boot}} = \frac{1}{B}\sum_{r=1}^{B} T^*_{n,b} - T_n.$$

We can also estimate the distribution $G_n(t) = P(T_n \leq t)$ of T_n. The bootstrap estimate of G_n is

$$\widehat{G}^*_n(t) = \frac{1}{B}\sum_{r=1}^{B} I\left\{T^*_{n,b} \leq t\right\}$$

where $I\{A\} = 1$ if A is true and 0 otherwise. A histogram of $T^*_{n,1}, \ldots, T^*_{n,B}$ can be regarded as an estimate of the density function of the distribution of T_n.

Under weak conditions on T_n, it can be shown that $\sup_t |\widehat{G}^*_n(t) - G_n(t)| \xrightarrow{p} 0$ as $n \to \infty$, i.e. the bootstrap is consistent. Similarly, it can be shown that the bootstrap variance estimate is consistent.

8.1 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. They vary in ease of calculation and accuracy. I will discuss two of them.

Normal Interval. If T_n is approximately normal, then we can construct an approximate confidence interval by:

$$T_n \pm z_{\alpha/2}\ \widehat{\text{se}}_{\text{boot}}$$

where $\text{se}_{\text{boot}} = \sqrt{\text{var}_{\text{boot}}(T_n)}$.

EXAMPLE 8.1 (The Mice Data.) *Let us return to the mice data. Suppose we are interested in the difference of the medians. We can compute the estimate and standard error as follows.*

```
x <- c(94, 197, 16, 38, 99, 141, 23)
y <- c(52, 104, 146, 10, 50, 31, 40, 27, 46 )
nx <- length(x)
ny <- length(y)
```

```
mx <- median(x)
my <- median(y)
delta <- mx - my
print(delta)
B <- 1000
delta.boot <- rep(0,1000)
for(i in 1:B){
     xstar <- sample(x,size=nx,replace=T)
     ystar <- sample(y,size=ny,replace=T)
     delta.boot[i] <- median(xstar) - median(ystar)
     }
se <- sqrt(var(delta.boot))
```

The estimate is 48 and the estimated standard error is 42. An approximate 95 per cent interval is $48 \pm 2(42) = (-36, 132)$.

Percentile Intervals. Let $T^*_{(1)}$ be the smallest bootstrap statistic, let $T^*_{(2)}$ be the second smallest bootstrap statistic, etc. Then $T^*_{(B\alpha)}$ is the α-percentile of the bootstrap values, that is, the value such that α of the statistics are smaller than it. Here it is understood that $B\alpha$ is rounded to an integer. The percentile interval defined by

$$\left(T^*_{(B\alpha/2)},\ T^*_{(B(1-\alpha/2))}\right).$$

The justification for this interval is given in the appendix.

EXAMPLE 8.2 *In the mouse data, we can get the percentile confidence interval as:*

```
quantile(delta.boot,c(.025,.975))
```

Another way is:

```
delta.boot <- sort(delta.boot)
i <- round(B*alpha/2)
delta.boot[i]
i <- round(B*(1-alpha/2))
delta.boot[i]
```

The interval is (-29,101).

The coverage of a bootstrap confidence interval is only approximately $1-\alpha$. There are more elaborate bootstrap confidence intervals that make this approximation more accurate. We won't go into details here.

8.2 Case Study 1

Here is an example that was one of the first used to illustrate the bootstrap by Bradley Efron, the inventor (discoverer?) of the bootstrap. The data are LSAT scores (for entrance to law school) and GPA.

```
LSAT   576   635   558   578   666   580   555   661
       651   605   653   575   545   572   594

GPA    3.39  3.30  2.81  3.03  3.44  3.07  3.00  3.43
       3.36  3.13  3.12  2.74  2.76  2.88  3.96
```

Each data point is of the form $X_i = (Y_i, Z_i)$ where $Y_i = \mathrm{LSAT}_i$ and $Z_i = \mathrm{GPA}_i$. The law school is interested in the correlation

$$\theta = \frac{\int (y-\mu_Y)(z-\mu_Z)dF}{\left[\int (y-\mu_Y)^2 dF \int (z-\mu_Z)^2 dF\right]^{1/2}}.$$

The plug-in estimate is the sample correlation

$$\widehat{\theta} = \frac{\sum_i (Y_i - \overline{Y})(Z_i - \overline{Z})}{\left[\sum_i (Y_i - \overline{Y})^2 \sum_i (Z_i - \overline{Z})^2\right]^{1/2}}.$$

Here is the R code for this problem. It contains a few good R tricks.

```
theta.fun <- function(y,z){
     ### this function computes the correlation coefficient
     mean.y <- mean(y)
     mean.z <- mean(z)
```

```
     s.y    <- sqrt(var(y))
     s.z    <- sqrt(var(z))
     top    <- sum((y-mean.y)*(z-mean.z))
     bottom <- sqrt( sum((y-mean.y)^2)*sum((z-mean.z)^2) )
     output <- top/bottom
     return(output)
     }

y <- c(576,635,558,578,666,580,555,661,651,605,653,575,545,572,594)
z <- c(3.39,3.30,2.81,3.03,3.44,3.07,3.00,3.43,3.36,3.13,
       3.12,2.74,2.76,2.88,2.96)


n          <- length(y)
theta.hat  <- theta.fun(y,z)
print(theta.hat)
B          <- 1000
theta.boot <- rep(0,B)
index <- 1:n
for(i in 1:B){
     j <- sample(index,replace=T)
     ystar <- y[j]
     zstar <- z[j]
     theta.boot[i] <- theta.fun(ystar,zstar)
     }
se <- sqrt(var(theta.boot))
print(se)

postscript("lsat.ps")
### remember to remove the postscript command
### if you are running R interactively
par(mfrow=c(2,1)) ### put several plots per page
                  ### 2 rows and 1 column of plots
plot(y,z,xlab="LSAT",ylab="GPA")
hist(theta.boot,nclass=20,xlab="Bootstrap Samples")
```

The estimated correlation is $\widehat{\theta} = .776$. The bootstrap gives $\widehat{se} = .137$. I used $B = 1000$. Figure 8.3.1 shows the data and a histogram of the bootstrap replications $\widehat{\theta}^{*1}, \ldots, \widehat{\theta}^{*B}$. This histogram is an approximation to the sampling distribution of $\widehat{\theta}$. We can also estimate the bias:

```
mean(theta.boot)-theta.hat
```

which yields $\widehat{\text{bias}} = -0.0005323285$. This is tiny compared to $\widehat{se}$ so the bias is not a concern. The Normal-based 95 per cent confidence interval is $.78 \pm 2(\widehat{se}) = (.51, 1.00)$ while the percentile interval is (.46,.96). In large samples, the two methods will show closer agreement.

8.3 Case Study II

This case study is borrowed from *An Introduction to the Bootstrap* by B. Efron and R. Tibshirani. When drug companies introduce new medications, they are sometimes requires to show *bioequivalence*. This means that the new drug is not substantially different than the current treatment. Here are data on eight subjects who used medical patches to infuse a hormone into the blood. Each subject received three treatments: placebo, old-patch, new-patch.

subject	placebo	old	new	old-placebo	new-old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

Let $Z = \text{old} - \text{placebo}$ and $Y = \text{new} - \text{old}$. The Food and Drug Administration (FDA) requirement for bioequivalence is that $|\theta| \leq .20$ where

$$\theta = \frac{E_F(Y)}{E_F(Z)}.$$

The estimate of θ is

$$\widehat{\theta} = \frac{\overline{Y}}{\overline{\overline{Z}}} = \frac{-452.3}{6342} = -.0713.$$

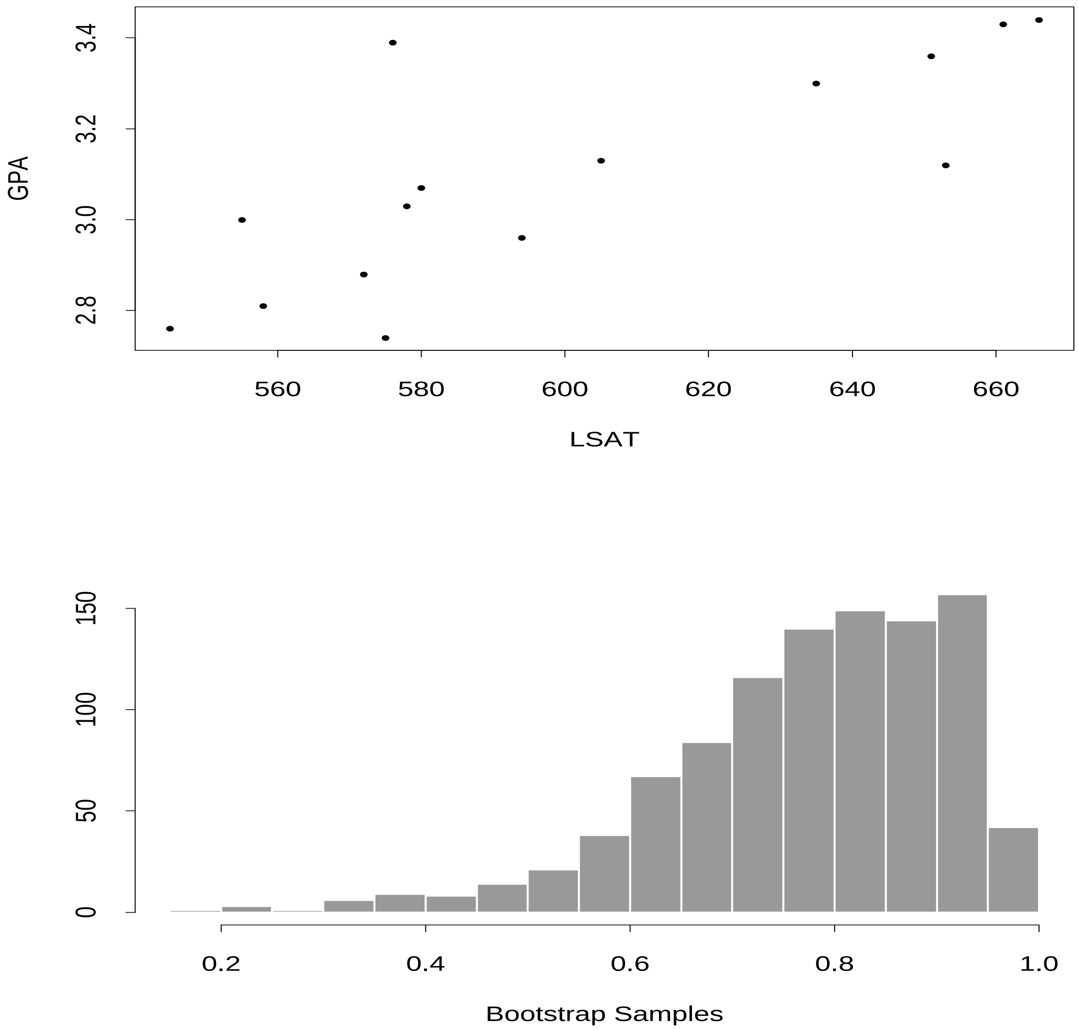

Figure 8.4.1. Law school data.

The bootstrap standard error is $\widehat{se} = .105$.

This estimator is the ratio of two averages. From experience, statisticians know that ratio estimators can sometimes be badly biased. The bootstrap estimate of the bias is $\overline{\theta}^* - \widehat{\theta} = -.0670 - (-.0713) = .0043$. This is small relative to $\widehat{se}$ so in this case, bias is not a problem.

To answer the bioequivalence question, let's compute a confidence interval.

```
## assume the bootstrap values are stored in a vector called theta.boot
alpha <- .05
j <- round(B*alpha/2)
k <- round(B*(1-(alpha/2)))
theta.boot <- sort(theta.boot)
lower <- theta.boot[j]
upper <- theta.boot[k]
print(lower)
print(upper)
### or do this:
quantile(theta.boot,.025)
quantile(theta.boot,.975)
```

From $B = 1000$ bootstrap replications we get the 95 per cent interval is [-.24,.15]. This is not quite contained in [-.20,.20] so at the 95 per cent level we have not demonstrated bioequivalence. Figure 8.5.1 shows the histogram of the bootstrap values.

8.4 Appendix: The Jackknife

The jackknife, due to Quenouille (1949), is a simple method for estimating the variance of a statistic. It is less computationally expensive than the bootstrap but is less general. Let $T_n = T(X_1, \ldots, X_n)$ be a statistic and $T_{(-i)}$ denote the statistic with the i^{th} observation removed. Let $\overline{T}_n = n^{-1}\sum_{i=1}^n T_{(-i)}$. The jackknife estimate of $\mathrm{var}(T_n)$ is

$$v_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^n (T_{(-i)} - \overline{T}_n)^2$$

and the jackknife estimate of the standard error is $\widehat{se}_{jack} = \sqrt{v_{jack}}$. This formula is very nonintuitive. It makes sense that the variance should involve

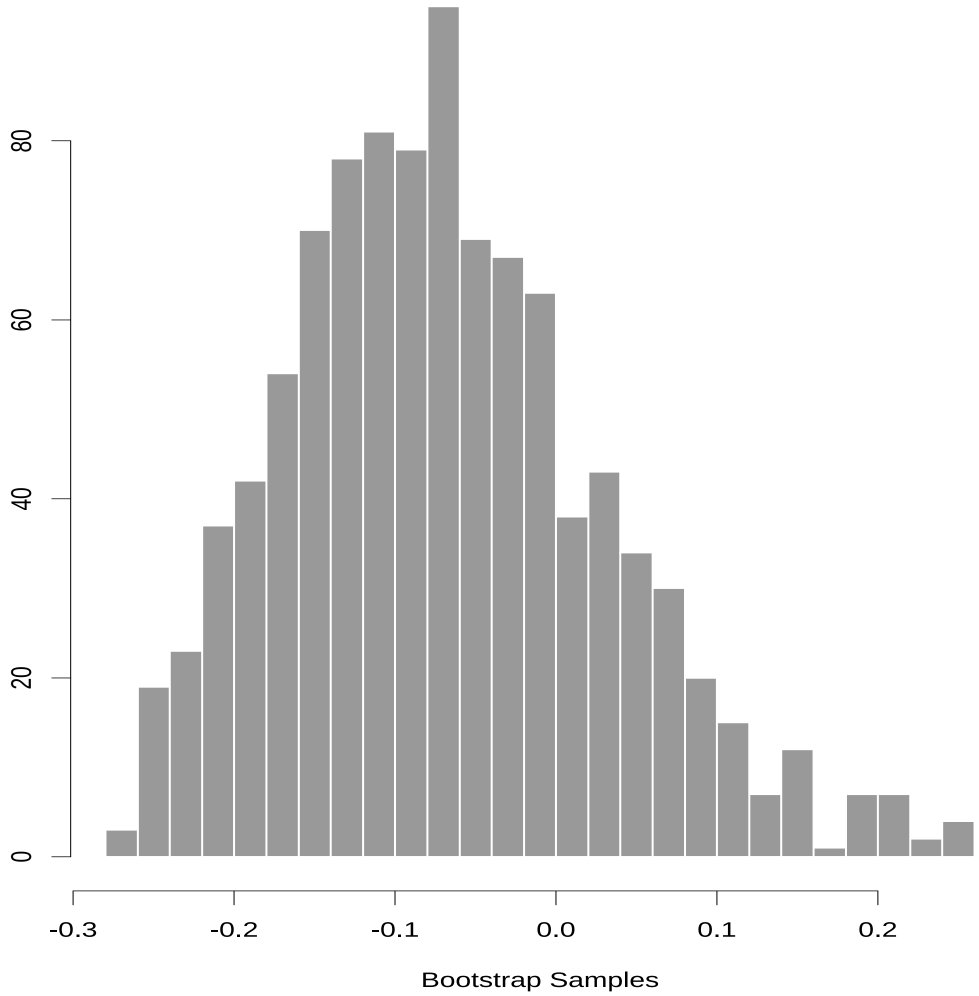

Figure 8.5.1. Patch data.

$\sum_{i=1}^{n}(T_{(-i)} - \overline{T}_n)^2$ but where does the $(n-1)/n$ term come from? The answer is related to the fact that T_n and $T_{(-i)}$ are based on different numbers of observations. To make the jackknife give the right answer in the special case $T_n = \overline{X}_n$, the factor $(n-1)/n$ turns out to be just what we need. See below for more discussion on this point.

Under suitable conditions on T, it can be shown that v_{jack} consistently estimates $\text{var}(T_n)$ in the sense that $v_{\text{jack}}/\text{var}(\text{T}_\text{n}) \xrightarrow{p} 1$.[8]

EXAMPLE 8.3 *Let $T_n = \overline{X}_n$. Some calculations show that $v_{\text{jack}} = S_n^2/n$ where S_n^2 is the sample variance, as expected.*

Once we have the estimate $\widehat{\theta}$ and the standard error $\widehat{se} = \sqrt{v_{jack}}$, we can form an approximate, Normal-based $1-\alpha$ confidence interval: $\widehat{\theta}_n \pm z_{\alpha/2}\,\widehat{se}$.

EXAMPLE 8.4 *Here is an example where the jackknife does* not *work. Suppose we want to estimate the median $\theta = F^{-1/2}$. The plug-in estimator is $\widehat{\theta} = \widehat{F}_n^{-1}(1/2)$. We defined $\widehat{F}_n^{-1}(1/2)$ to be smallest number t such that $\widehat{F}_n(t) \geq 1/2$. Let $X_{(1)}, \ldots, X_{(n)}$ denote the data ordered from smallest to largest. Then, according to our definition, $\widehat{\theta}_n = X_{(n/2)}$ if n is even and $\widehat{\theta}_n = X_{((n+1)/2)}$ if n is odd. But it turns out that the estimated standard error from the jackknife does not give a consistent estimate of the true standard error. The reason is that the jackknife only works when T_n is "smooth." The median is not a smooth functional. To see this, consider taking the smallest data point and then increasing it. At first this will not affect the median. Eventually, the median will jump and then it will stay constant again. So the median does not change smoothly as we move data points around. The bootstrap is superior because it does provide consistent estimates of the standard error of the median (and other unsmooth functionals.)*

The jackknife can also be used to estimate $\text{bias}(T_n) = E(T_n) - \theta$. The jackknife bias estimate is defined by

$$b_{jack} = (n-1)(\overline{T}_n - T_n). \tag{4}$$

The *bias-corrected estimate* is defined to be $T_{\text{jack}} = T_n - b$.

Now we give some explanation about the form of the jackknife bias estimate. Define the *pseudo-values*

$$\widetilde{T}_i = nT_n - (n-1)T_{(-i)}.$$

[8] Strictly speaking, the jackknife estimates the asymptotic variance.

T_{jack} can then be written as

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{T}_i.$$

Also, we can write

$$v_{\text{jack}} = \frac{\widetilde{s}^2}{n} \tag{5}$$

where

$$\widetilde{s}^2 = \frac{\sum_{i=1}^{n}\left(\widetilde{T}_i - \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_i\right)^2}{n-1}$$

is the sample variance of the pseudo-values. (One can use n instead of $n-1$ in the denominator if one prefers.)

If a_n and b_n are sequences, we write $a_n = O(b_n)$ to mean that $|a_n/b_n|$ is bounded for all large n. For many statistics, it turns out that

$$\text{bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

for some a and b. Therefore,

$$\text{bias}(T_{(-i)}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

The same expression holds for bias$(\overline{T}_n)$. Hence,

$$\begin{aligned} E(b) &= (n-1)(\text{bias}(\overline{T}_n) - \text{bias}(T_n)) \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\ &= \text{bias}(T_n) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

which shows that b_{jack} estimates the bias up to order $O(n^{-2})$. Also,

$$\text{bias}(T_{\text{jack}}) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

so the bias of T_{jack} is an order of magnitude smaller than that of T_n.

8.5 Appendix: Bootstrap Percentile Confidence Intervals

Suppose there exists a monotone transformation $U = m(T)$ such that $U \sim N(\phi, c^2)$ where $\phi = m(\theta)$. We do not suppose we know the transformation, only that one exist. Let $U_b^* = m(T_b^*)$. Note that $U^*_{(B\alpha/2)} = m(T^*_{(B\alpha/2)})$ since a monotone transformation preserves quantiles. Since, $U \sim N(\phi, c^2)$, the $\alpha/2$ quantile of U is $\phi - z_{\alpha/2}c$. And, $U^*_{(B\alpha/2)}$ is the $\alpha/2$ quantile so $U^*_{(B\alpha/2)} = \phi - z_{\alpha/2}c$. Similar comments apply to the upper quantiles. Therefore,

$$\begin{aligned}
\Pr\{T^*_{B\alpha/2} \leq \theta \leq T^*_{B(1-\alpha/2)}\} &= \Pr\{m(T^*_{B\alpha/2}) \leq m(\theta) \leq m(T^*_{B(1-\alpha/2)})\} \\
&= \Pr\{U^*_{B\alpha/2} \leq \phi \leq U^*_{B(1-\alpha/2)}\} \\
&\approx \Pr\{U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}\} \\
&= \Pr\{-z_{\alpha/2} \leq \frac{\phi - U}{c} \leq z_{\alpha/2}\} \\
&= 1 - \alpha.
\end{aligned}$$